Data-driven selection and parameter estimation for DNA methylation mathematical models

Karen Larson[a], Loukas Zagkos[b], Mark Mc Auley[c], Jason Roberts[b], Nikos I. Kavallaris[b], Anastasios Matzavinos[a,*]

[a] *Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912, USA*

[b] *Department of Mathematics, School Of Science and Engineering, University of Chester, Thornton Science Park, Pool Lane, Ince Chester CH2 4NU, UK*

[c] *Department of Chemical Engineering, School Of Science and Engineering, University of Chester, Thornton Science Park, Pool Lane, Ince,Chester CH2 4NU, UK*

Abstract

Epigenetics is coming to the fore as a key process which underpins health. In particular emerging experimental evidence has associated alterations to DNA methylation status with healthspan and aging. Mammalian DNA methylation status is maintained by an intricate array of biochemical and molecular processes. It can be argued changes to these fundamental cellular processes ultimately drive the formation of aberrant DNA methylation patterns, which are a hallmark of diseases, such as cancer, Alzheimer's disease and cardiovascular disease. In recent years mathematical models have been used as effective tools to help advance our understanding of the dynamics which underpin DNA methylation. In this paper we present linear and nonlinear models which encapsulate the dynamics of the molecular mechanisms which define DNA methylation. Applying a recently developed Bayesian algorithm for parameter estimation and model selection, we are able to estimate distributions of parameters which include nominal parameter values. Using limited noisy observations, the method also identified which methylation model the observations originated from, signaling that our method has practical applications in identifying what models best match the biological data for DNA methylation.

*Corresponding author

Email address: matzavinos@brown.edu (Anastasios Matzavinos)

Keywords: DNA methylation, model selection, parameter estimation, gene promoter, CpG dyads

1. Introduction

DNA methylation has a pivotal epigenetic role to play during embryonic development [1]. The covalent bonding of methyl groups to DNA serves to regulate gene expression during this period and is a process which culminates with the formation of tissue specific methlylation patterns. However, during ageing mammalian methylation patterns change. Ageing is synonymous with genome wide hypomethylation, whilst, paradoxically, it is associated with regional increases in DNA methylation, most notably at the promoter region of a diverse array of genes [2].

Intriguingly, several disease processes display similar characteristics. Specifically, cancers invariably show global hypomethylation and gene specific hypermethylation, while autoimmune diseases routinely exhibit hypomethylation both globally and on specific genes [3]. Moreover, changes to genomic methylation patterns with age have a burgeoning role to play in cardiovascular disease [4], Alzheimer's disease [5], and osteoporosis/osteoarthritis [6]. Thus, it is clear the dysregulation of this fundamental epigenetic process is vital to a variety of age related pathologies and potentially ageing. In order to identify why an increase in age results in aberrant DNA methylation, it is necessary to understand the molecular mechanisms which govern this biochemical system [7, 8]. Additionally, it is imperative to appreciate how the dynamics of this molecular system change with age.

DNA methylation occurs in mammals primarily at CpG dyads; more specifically, the methyl group is attached to the fifth carbon of the cytosine at the CpG site (Cytosine - Guanine dinucleotide sequence separated by a phosphate group). Within the vertebrate genome global methylation can be defined by CpG islands (CGIs). These are genomic regions which comprise 1000 base pairs and consist of high levels of G+C base levels. In addition, they are characterized by a deficiency in DNA methylation [9]. Although CGI are scantly decorated throughout the genome, their biological imperative has been coming to the fore in recent years. Chiefly, CGIs are sites of transcriptional initiation and thus act as promoters in mammalian genomes. Consequently, any change in the methylation status of a CGI will potentially effect gene transcription, and this is exactly what happens as hypermethylation of CGIs are routinely correlated with the transcriptional silencing of

gene promoters, a phenomenon which is often a feature of diseases such as cancer [10]. Besides, increasing age has been correlated with the hypermethylation of a wide variety of gene promoters belonging to genes which have been associated with ageing [11]. Consequently, it is clear from the above discussion ageing has a profound effect on the dynamics of DNA methylation and age related changes to processes which control the reactions which govern DNA methylation ultimately drive the formation of aberrant DNA methylation.

This biological system remains to be fully delineated; what is known is that it is characterized by the activities of several enzymes [12]. The enzymes operate as follows: post replicatively, new CpG dinucleotides are attached to the complementary strand of the daughter cells, which are unmethylated. DNA methyltransferase (Dnmt1) then uses S-Adenosyl methionine as a substrate to transfer methyl groups to the DNA molecule [13]. As Dnmt1 preferentially acts on hemimethylated DNA it is thought to be chiefly a maintenance enzyme [14]. Therefore, other enzymes are a necessity for *de novo* DNA methylation. Current thinking suggests Dnmt3a and Dnmt3b are the enzymes which perform this task. Enzymatic maintenance and *de novo* methylation reactions are in turn counterbalanced by passive and active demethylation [15]. Passive demethylation usually occurs as a result of replication and DNA methylation levels can decrease after several rounds of this process [16]. On the other hand it is suggested active methylation requires Ten-Eleven Translocation (TET) dioxygenases, which oxidize the methyl groups of cytosine; a process which eventually results in the reincorporation of an unmethylated cytosine into DNA [17]. As a result the maintenance of DNA methylation levels can be viewed as a subtle balancing act between maintenance/*de novo* methylation and passive/active demethylation.

In recent years mathematical models have been used as effective tools to help advance our understanding of the dynamics which underpin DNA methylation – reviewed in [18]. In the current work we present linear and nonlinear mathematical models which encapsulate the molecular mechanisms which define DNA methylation [19]. In addition, we use our recently developed Bayesian algorithm for estimating the parameters of a model and, furthermore, select the model that best fits given DNA methylation data. The Bayesian parameter estimation allows us to leverage prior knowledge of the methylation rates to guide the search in the parameter space. To test the viability of using parallel transitional Markov chain Monte Carlo (TMCMC)

for the DNA methylation problem, parameters are estimated using noisy model observations generated from two potential models. Furthermore, the sampling algorithm used allows for model selection without any additional computational resources. With this in mind, we will use the algorithm to identify which model best matches noisy model observations and will test which model is most biologically feasible. The model selected as "best" is the one that most closely fits the biological system and discovers correlations that match with the underlying biological mechanisms.

2. Models and Methods

2.1. DNA methylation models

Based on the approach in [20], three types of population are considered; unmethylated CpG dyads, the total number of which is denoted as $x_1(t)$, hemimethylated CpG dyads, $x_2(t)$, and methylated CpG dyads, as $x_3(t)$, see Fig. 1. An unmethylated CpG dyad is a CpG dyad with none of the two CpG sites methylated. Similarly, a hemimethylated CpG dyad has only one methylated CpG site and the opposing unmethylated and a methylated CpG dyad has both opposing sites methylated. The methylation enzymes DNMT1, DNMT3a and DNMT3b, demethylation enzymes TET family and DNA replication are responsible for the transitions between the possible states of CpG dyads. The methylation rates of unmethylated CpG dyads and hemimethylated CpG dyads are k_1 and k_2, respectively. In addition, the demethylation rates of hemimethylated and methylated CpG dyads are k_3 and k_4, respectively. D denotes the rate of cell division, see Fig. 2.

Inspired by [21] the mechanism behind DNA division is described as follows. Unmethylated DNA strands bond with the parental strands during DNA replication. Therefore, all parental methylated CpG dyads form hemimethylated CpG dyads in the daughter cells [22]. The hemimethylated CpG dyads in the parental cell either become unmethylated, or remains hemimethylated in the daughter cells. In this case it is assumed that half of the parental hemimethylated CpG dyads become unmethylated in the daughter cells and the other half remain hemimethylated. Unmethylated dyads remain unmethylated.

The above biological mechanisms can be translated in a set of ordinary differential equations (ODEs) as follows, see [19],

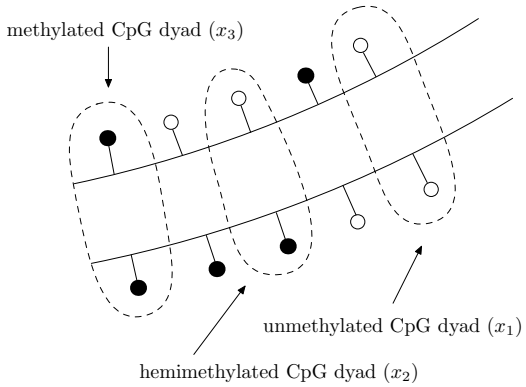


Figure 1: The three different states of a CpG dyad; unmethylated (x_1), hemimethylated (x_2) and methylated (x_3) CpG dyads. A white circle denotes an unmethylated CpG site whereas a black circle represents a methylated CpG site. An unmethylated (x_1) CpG dyad consists of two unmethylated opposite CpG sites, a hemimethylated dyad (x_2) has only one of the two sites methylated and a methylated dyad (x_3) has both opposing sites methylated.

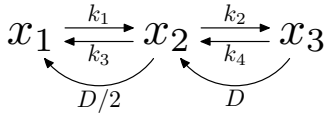


Figure 2: The diagram for the methylation rates between x_1, x_2 and x_3.

$$\frac{dx_1(t)}{dt} = -k_1 x_1(t) + \left(k_3 + \frac{1}{2}D\right)x_2(t) \tag{1}$$

$$\frac{dx_2(t)}{dt} = k_1 x_1(t) - \left(k_2 + k_3 + \frac{1}{2}D\right)x_2(t) + \left(k_4 + D\right)x_3(t) \tag{2}$$

$$\frac{dx_3(t)}{dt} = k_2 x_2(t) - \left(k_4 + D\right)x_3(t). \tag{3}$$

Gene promoters are regions of interest in terms of DNA methylation levels. Monitoring the evolution of the populations of CpG sites in gene promoters, dictates that the total number of CpG dyads has to be constant. Therefore, it was considered that $x_1(t)+x_2(t)+x_3(t) = C$, with $C > 0$. Substituting the above equation into the set of equations (1)-(3), we deduce the equivalent non-homogeneous system

$$\frac{dx_1(t)}{dt} = -\left(k_1 + k_3 + \frac{1}{2}D\right)x_1(t) - \left(k_3 + \frac{1}{2}D\right)x_3(t) + C\left(k_3 + \frac{1}{2}D\right) \quad (4)$$

$$\frac{dx_3(t)}{dt} = -k_2 x_1(t) - \left(k_2 + k_4 + D\right)x_3(t) + Ck_2 \quad (5)$$

$$x_2(t) = C - x_1(t) - x_3(t). \quad (6)$$

Depending on the specific type of tissue, methylation levels in gene promoters may significantly vary [23]. There are two different patterns observed in the DNA methylation levels in gene promoters; hypomethylated and hypermethylated [24]. In a hypomethylated location, the vast majority of the CpG sites are unmethylated. On the contrary, a hypermethylated region consists mainly of methylated CpG sites. Thus, it is necessary to introduce some nonlinear terms to obtain the observed behaviour, namely, the bistable state of gene promoters. A key question is how to determine the most appropriate nonlinear model that will give the expected behaviour with respect to the data collected.

To account for a potential transition between the two states it is necessary to appreciate the following biological arguments. If a scenario exists whereby there is an abundance of unmethylated CpG dyads (x_1) and the gene promoter is hypomethylated, it is reasonable to assume that with time unmethylated CpG dyads will become methylated. Biologically this could happen as a result of fluctuating levels of DNMT3a and DNMT3b (denoted in the model by an increase in k_1 rate). As the number of unmethylated CpG dyads (x_1) drops, the methylation rate k_1 increases. While the level of unmethylated dyads decreases, then the number of hemimethylated dyads (x_2) increases. This can be interpreted as k_1 being a decreasing function of $x_1(t)$ or an increasing function of $x_2(t)$. It remains unknown if the transition between the two different states is due to an increase in the *de novo* methylation enzymes (DNMT3a and DNMT3b) or whether it is due to a decrease in the demethylation enzymes (the TET protein family). To account for the latter, it can be assumed that as the number of unmethylated CpG dyads (x_1) decreases, the demethylation rate k_3 drops, due to a change in TET enzyme activity and consequently the number of hemimethylated CpG dyads increases. This can be described by denoting the rate k_3 as an increasing function of $x_1(t)$ or a decreasing function of $x_2(t)$. A similar premise can be suggested for the transition rates k_2 and k_4. A significant rise in methylated

CpG dyads can be as a result of an increase in DNMT1 maintenance levels or due to a decrease in the TET enzymes, namely either a k_2 increase or a k_4 drop. Thus it is logical that k_2 can be a decreasing function of x_2 or an increasing function of x_3 and k_4 an increasing function of x_2 or a decreasing function of x_3.

Following the biological assumptions made above, there are two plausible approaches of representing the transition rates as functions of the CpG dyads. These expressions can be considered either in terms of x_2 or in terms of x_1 and x_3, as follows.

$$k_1(x_2) = k_{11} + k_{12}x_2^{\gamma_1}(\nearrow), \tag{7}$$
$$k_2(x_2) = k_{21} + k_{22}x_2^{\gamma_2}(\searrow), \tag{8}$$
$$k_3(x_2) = k_{31} + k_{32}x_2^{\gamma_3}(\searrow), \tag{9}$$
$$k_4(x_2) = k_{41} + k_{42}x_2^{\gamma_4}(\nearrow), \tag{10}$$

or

$$k_1(x_1) = k_{11} + k_{12}x_1^{\gamma_1}(\searrow), \tag{11}$$
$$k_2(x_3) = k_{21} + k_{22}x_3^{\gamma_2}(\nearrow), \tag{12}$$
$$k_3(x_1) = k_{31} + k_{32}x_1^{\gamma_3}(\nearrow), \tag{13}$$
$$k_4(x_3) = k_{41} + k_{42}x_3^{\gamma_4}(\searrow), \tag{14}$$

where $k_j(x_i) > 0$ and $\gamma_i \in \mathbb{R}$.Here we assume $\gamma_i = 2$, since these reactions are akin to second order kinetics which are common in biochemical systems. The arrow next to each formula denotes an increasing or decreasing transition function of the populations x_i, as biology dictates. In our previous work [19], we selected methylation rates as functions of x_1 and x_3. Therefore, the following system is obtained

$$\frac{dx_1(t)}{dt} = -A_1(x_1(t))x_1(t) - A_2(x_1(t))x_3(t) + A_3(x_1(t))C \tag{15}$$
$$\frac{dx_3(t)}{dt} = -B_1(x_3(t))x_1(t) - B_2(x_3(t))x_3(t) + B_3(x_3(t))C \tag{16}$$
$$x_2(t) = C - x_1(t) - x_3(t), \tag{17}$$

where

$$
\begin{aligned}
A_1(x_1(t)) &= k_{11} - k_{12}x_1^2(t) + k_{31} + k_{32}x_1^2(t) + \frac{1}{2}D, \\
A_2(x_1(t)) &= k_{31} + k_{32}x_1^2(t) + \frac{1}{2}D, \\
A_3(x_1(t)) &= k_{31} + k_{32}x_1^2(t) + \frac{1}{2}D, \\
B_1(x_3(t)) &= k_{21} + k_{22}x_3^2(t), \\
B_2(x_3(t)) &= k_{21} + k_{22}x_3^2(t) + k_{41} - k_{42}x_3^2(t) + D, \\
B_3(x_3(t)) &= k_{21} + k_{22}x_3^2(t),
\end{aligned}
$$

see [19]. The results of our model corroborate experimental work which has investigated the epigenetic nature of gene promoters [25]. Moreover, sensitivity analysis was able to suggest which parameters were vulnerable to small perturbations. However, given the uncertainty which surrounds these parameter values generally and the lack of quantitative biological information, it is necessary to consolidate our findings. One approach to these problems is through utilizing Bayesian inference.

2.2. Bayesian Uncertainty Quantification

It is reasonable to assume that in reality DNA methylation will not exactly match any model and measured data will be noisy. Statistical inference is included to infer information from observations. Bayesian uncertainty quantification (UQ) assumes that parameters are random variables with unknown distributions and leverages prior information, knowledge, and experience to inform searches about distributions of unknown parameters.

2.2.1. Model Parameter Estimation

In this context, the parameters of interest $\underline{\theta}$ are inputs into a DNA methylation model M that predicts output quantities of interest $g(\underline{\theta}|M) \in \mathbb{R}^m$, e.g. the number of unmethylated, hemimethylated and methylated CpG dyads x_1, x_2, and x_3. As the model cannot exactly represent physical, observed quantities $\underline{D}$ due to various errors (e.g. measurement, computational or modelling), the Bayesian context needs an explicit expression relating the model outputs to the noisy observation data. One possible perturbation is that the observed data $\underline{D}$ are generated according to the model prediction equation:

$$\underline{D} = g(\underline{\theta}|M) + \underline{e}, \tag{18}$$

where $g(\underline{\theta}|M)$ are the model predictions for a given model inputs $\underline{\theta} \in \mathbb{R}^n$ and $\underline{e}$ is the prediction error. The posterior distribution for the parameters given the observed data is given by Bayes' Theorem as:

$$p(\underline{\theta}|\underline{D}, M) = \frac{p(\underline{D}|\underline{\theta}, M)\pi(\underline{\theta}|M)}{\rho(\underline{D}|M)}, \tag{19}$$

in terms of the prior distribution on the parameters $\pi(\underline{\theta}|M)$, likelihood $p(\underline{D}|\underline{\theta}, M)$, and evidence $\rho(\underline{D}|M)$ of the model class, given by the multi-dimensional integral

$$\rho(\underline{D}|M) = \int_{\mathbb{R}^n} p(\underline{D}|\underline{\theta}, M)\pi(\underline{\theta}|M)d\underline{\theta}.$$

When M is one particular model in a parameterized class of models, the evidence $\rho(\underline{D}|M)$ serves as a measure of how well the model matches the data and serves as one method for model selection [26, 27].

Using the prediction error equation (18) and assuming that the prediction errors $\underline{e}$ are Gaussian distributed with mean 0 and covariance matrix Σ, the observed data $\underline{D}$ will also be normally distributed. Thus, the likelihood $p(\underline{D}|\underline{\theta}, M)$ is given by

$$p(\underline{D}|\underline{\theta}, M) = \frac{|\Sigma(\underline{\theta})|^{-1/2}}{(2\pi)^{m/2}} \exp\Big[-\frac{1}{2}J(\underline{\theta}, \underline{D}|M)\Big]$$

where

$$J(\underline{\theta}, \underline{D}|M) = [\underline{D} - g(\underline{\theta}|M)]^T \Sigma^{-1}(\underline{\theta})[\underline{D} - g(\underline{\theta}|M)]$$

is the weighted measure of fit between the model predictions and measured data, $|\cdot|$ denotes the determinant, and the parameter set $\underline{\theta}$ is augmented to include parameters that are involved with the structure of the covariance matrix Σ.

2.2.2. Model Selection

The Bayesian uncertainty quantification framework can be extended to not only estimate distributions of parameters, but also compare the plausibility of different models based upon the available data. In this case, there

is a family $\mathcal{M} = \{M_i, i = 1, \ldots, \kappa\}$ of κ alternative model classes. Each of these models in our context refers to a different expression for k_j, which can depend either on the hemimethylated CpG dyads x_2 or on the unmethylated and methylated CpG dyads x_1 and x_3.

Similar to parameter estimation, we assume a prior distribution on the different model classes $Pr(M_i)$, which corresponds to the probability of selecting model M_i from the family $\mathcal{M}$. Using Bayes' rule, the posterior probability for model M_i is given as:

$$Pr(M_i|\underline{D}) = \frac{\rho(\underline{D}|M_i)Pr(M_i)}{p(\underline{D}|\mathcal{M})}$$

where $Pr(M_i|\underline{D})$ is the posterior distribution for model class M_i, $\rho(\underline{D}|M_i)$ is the evidence for model, and $p(\underline{D}|\mathcal{M}) = \sum_{i=1}^{\kappa} \rho(\underline{D}|M_i)$ is a normalization constant. If the prior on models is uniform, then the posterior distribution $Pr(M_i|\underline{D})$ for each model is directly proportional to the evidence $\rho(\underline{D}|M_i)$. Therefore model selection is free when the evidence has already been calculated in parameter estimation.

2.2.3. TMCMC Method

As the parameter distributions are often unknown, we need to use a method to approximately sample from them to estimate the underlying posterior distribution. One such sampling method is transitional Markov chain Monte Carlo (TMCMC), which benefits from its ability to run a large number of Markov chains in parallel, alleviating some of the computational bottlenecks that often occur with sampling methods.

The TMCMC algorithm used by the highly efficient task sharing framework Π4U [28, 29, 30] slowly transitions from the prior distribution to the target distribution (the posterior $p(\underline{\theta}|\underline{D}, M)$) by constructing a series of intermediate distribution functions:

$$f_j(\underline{\theta}) \sim [p(\underline{D}|\underline{\theta}, M)]^{q_j} \cdot \pi(\underline{\theta}|M), \;\; j = 0, \ldots, \lambda$$
$$0 = q_0 < q_1 < \ldots < q_\lambda = 1.$$

The TMCMC algorithm is summarized in Algorithm 1. Initially, N_0 samples $\underline{\theta}_{0,k}$ are taken from the prior distribution $f_0(\underline{\theta}) = \pi(\underline{\theta}|M)$. For each stage j of the algorithm, the current samples are evaluated by computing the plausibility weights $w(\underline{\theta}_{j,k})$ as

$$w(\underline{\theta}_{j,k}) = \frac{f_{j+1}(\underline{\theta}_{j,k})}{f_j(\underline{\theta}_{j,k})} = [p(\underline{D}|\underline{\theta}_{j,k}, M)]^{q_{j+1}-q_j}.$$

Since the q_j are selected to be monotonically increasing, the plausibility weights for parameter $\underline{\theta}_{j,k}$ are higher for those with a larger likelihood, i.e. ones that generate the observed data given the parameters and models. Secondly, the q_j's determine how smoothly the prior distribution transitions to the posterior distributions, with small increments yielding smoother updates, but a more computationally intensive algorithm. In order to balance computational efficiency and smooth transitions between intermediate distributions, recent literature suggests that q_{j+1} should be taken so that the covariance of the plausibility weights at stage j is smaller than a tolerance covariance value, often 1.0 [31, 28, 29, 30].

Algorithm 1 TMCMC

1: **procedure** TMCMC Ref. [28]
2: BEGIN, SET $j = 0, q_0 = 0$
3: **Generate** $\{\underline{\theta}_{0,k}, k = 1, \ldots, N_0\}$ from prior $f_0(\underline{\theta}) = \pi(\underline{\theta}|M)$ and compute likelihood $p(\underline{D}|\underline{\theta}_{0,k}, M)$ for each sample.
4: loop:
5: **WHILE** $q_{j+1} \leq 1$ **DO:**
6: **Analyze** samples $\{\underline{\theta}_{j,k}, k = 1, \ldots, N_j\}$ to determine q_{j+1}, weights $\overline{w}(\underline{\theta}_{j,k})$, covariance Σ_j, and estimator S_j of $\mathbb{E}[w(\underline{\theta}_{j,k})]$.
7: **Resample** based on samples available in stage j in order to generate samples for stage $j + 1$ and compute likelihood $p(\underline{D}|\underline{\theta}_{j+1,k}, M)$ for each.
8: **if** $q_{j+1} > 1$ **then**
9: BREAK,
10: **else**
11: $j = j + 1$
12: **goto** loop.
13: **end**
14: **END**

Next, the algorithm computes the average S_j of the plausibility weights, the normalized plausibility weights, the scaled covariance Σ_j of the samples

$\underline{\theta}_{j,k}$, which are used to produce the next generation of samples $\underline{\theta}_{j+1,k}$:

$$S_j = \frac{1}{N_j} \sum_{k=1}^{N_j} w(\underline{\theta}_{j,k})$$

$$\bar{w}(\underline{\theta}_{j,k}) / \sum_{k=1}^{N_j} w(\underline{\theta}_{j,k}) = w(\underline{\theta}_{j,k})/(N_j S_j)$$

$$\Sigma_j = b^2 \sum_{k=1}^{N_j} \bar{w}(\underline{\theta}_{j,k})[\underline{\theta}_{j,k} - \underline{\mu}_j][\underline{\theta}_{j,k} - \underline{\mu}_j]^T.$$

Σ_j is calculated using the sample mean $\underline{\mu}_j$ and a scaling factor b, usually taken to be 0.2 [31, 28, 29, 30].

The algorithm then generates N_{j+1} samples $\hat{\underline{\theta}}_{j+1,k}$ by randomly selecting from the previous generations of samples $\{\underline{\theta}_{j,k}\}$ such that $\hat{\underline{\theta}}_{j+1,\ell} = \underline{\theta}_{j,k}$ with probability $\overline{w}(\underline{\theta}_{j,k})$. These samples are selected independently at random, so any parameter can be selected multiple times. Let $n_{j+1,k}$ be the number of times $\underline{\theta}_{j,k}$ is selected. Each unique sample is used as the starting point of an independent Markov chain of length $n_{j+1,k}$ generated using the Metropolis algorithm [32] with target distribution f_j and a Gaussian proposal distribution with covariance Σ_j centered at the current value. The Metropolis algorithm for each of our Markov chains yields N_{j+1} total samples $\underline{\theta}_{j+1,k}$. Finally, the algorithm either moves forward to generation $j+1$ or terminates if $q_{j+1} > 1$.

3. Results and Discussion

We apply Π4U to the nonlinear DNA methylation model described earlier. The four-stage Runge-Kutta method was used to generate the model outputs g in the model prediction equation. The model outputs considered are the time-series evolution for x_1, x_2, and x_3 from $t = 0$ to $t = 10$, with step size $\Delta t = 0.0001$ and every hundredth step recorded, resulting in 300 sample points. Measured data are simulated by computing all outputs using a reference model and corrupting each output by Gaussian noise as

$$D_k = \xi_k + \sigma\epsilon_k$$

where D_k is the observation data from the k^{th} position of the vector, ξ_k is the k^{th} model output, ϵ_k is a zero-mean, unit-variance Gaussian variable,

and σ is the level of the noise. The reference model is selected to correspond to some nominal values of the model parameters. In order for the signal-to-noise ratio to be high enough for meaningful estimation, we choose σ to be a fraction $\sigma = 0.01\alpha$ of the standard deviation α of all model outputs. The model prediction error covariance Σ is assumed to be a diagonal matrix $\Sigma = \sigma I$ whose nonzero entries all have the same magnitude σ.

In the following results, we estimate parameters via the generation of 10^4 samples from the posterior for the DNA methylation model. For ease of comparison, numerical results are computed in terms of the rescaled parameters ($\theta_{k_{31}}$, $\theta_{k_{32}}$, θ_γ, σ/α), given by $\theta_{k_{31}} = k_{31}/k_{31_0}$, i.e., the ratio between the estimated value and the nominal value. The prior is assumed uniform on $[-4, 4]\times[-4, 4]\times[-4, 4]\times[0, 0.05]$ in the scaled parameter space. We consider two formats for the non-linear term k_3: $k_3(x_1) = k_{31} + k_{32}x_1^\gamma$, an increasing function of x_1, and $k_3(x_2) = k_{31} + k_{32}x_2^\gamma$, a decreasing function of x_2. We take the following parameter values for the nominal parameter values $k_{31_0} = 1$, $k_{32_0} = 0.01$, $\gamma_0 = 2$, and $\sigma = 0.01\alpha$ for $k_3(x_1)$ and $k_{31_0} = 100$, $k_{32_0} = -0.01$, $\gamma_0 = 2$, and $\sigma = 0.01\alpha$ for $k_3(x_2)$.

3.1. Parameter Estimation

In this case, we have two sets of reference data. The first, denoted O_1, comes from the time history generated by the model M_1 where the expression k_3 depends on x_1 and the second, denoted O_2, comes from the model M_2 where the expression k_3 depends on x_2. In both cases, $k_1 = 0.012$, $k_2 = 99$, $k_4 = 0.08$ and the other parameters are as described above.

The results using $k_3(x_1)$ are displayed in Figure 3, which show a strong negative correlation with k_{31} and k_{32}, which determine the hemimethylation rate of CpG dyads (x_2 to x_1). This makes intuitive sense, as an increase in k_{31} should correspond with a decrease in k_{32} to match the dynamics. Correspondingly, both k_{32} and γ have a strong negative correlation.

The recovered scaled mean parameter values are in Table 1, recovering values $(k_{31}, k_{32}, \gamma, \sigma/\alpha) = (0.961,\ 0.0105,\ 1.998,\ 0.010)$. To quantify the degree of uncertainty for each parameter's posterior distribution, we compute the coefficient of variation, defined as the ratio of its standard deviation to its mean (denoting the results $u_{k_{31}}$, $u_{k_{32}}$, u_γ, $u_{\sigma/\alpha}$). In all cases, the parameter values are recovered within one standard deviation of the nominal values.

The results using $k_3(x_2)$ are displayed in Figure 4, which displays different correlations than the $k_3(x_1)$ case. Here, the k_{31} parameter is found almost exactly, while k_{32} and γ have a range of negative and positive values that are

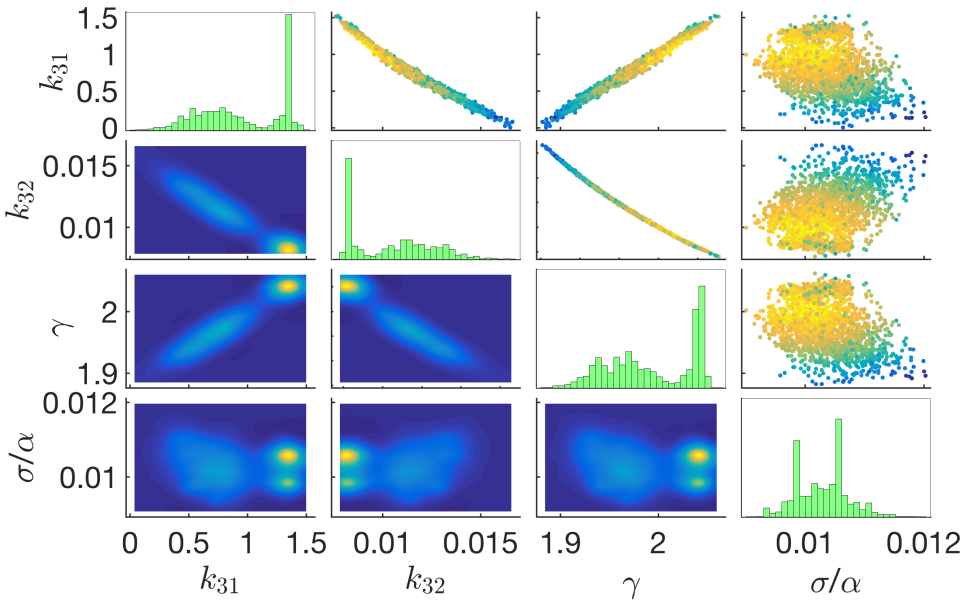


Figure 3: Parameter estimation results using reference data from the model where $k_3 = k_{31} + k_{32}x_1^\gamma$ with a time history from $T = 0$ to $T = 10$. The nominal parameter values used were $k_{31} = 1$, $k_{32} = 0.01$, $\gamma = 2$, and noise level $\sigma = 0.01\alpha$. The model used in parameter estimation is $k_3 = k_{31} + k_{32}x_1^\gamma$. Histograms for each parameter are displayed along the main diagonal of the figure. Sub-figures below the diagonal show the marginal joint density functions for each pair of parameters, while sub-figures above the diagonal show the samples used in the final stage of TMCMC. Colors correspond to probabilities, with yellow likely and blue unlikely.

found. This is likely due to small changes in k_{32} and γ not greatly effecting the dynamics of $k_3(x_2)$.

The recovered scaled mean parameter values are in Table 1, recovering values $(k_{31}, k_{32}, \gamma, \sigma/\alpha)$ = (99.1, -0.014, 0.576, 0.010). In this case, the parameters are all recovered within two standard deviations of the mean, due to the wide smear in the recovered parameter values.

3.2. Model Selection

Next, our aim was to identify which model generated which reference data set. To do this, we used the reference data set for $k_3(x_2)$ and use the $k_3(x_1)$ model to recover parameters; correspondingly the $k_3(x_2)$ model also uses the $k_3(x_1)$ reference data to perform parameter estimation. The results for these two experiments are displayed in Figures 5a and 5b, respectively. Since $k_3(x_2)$

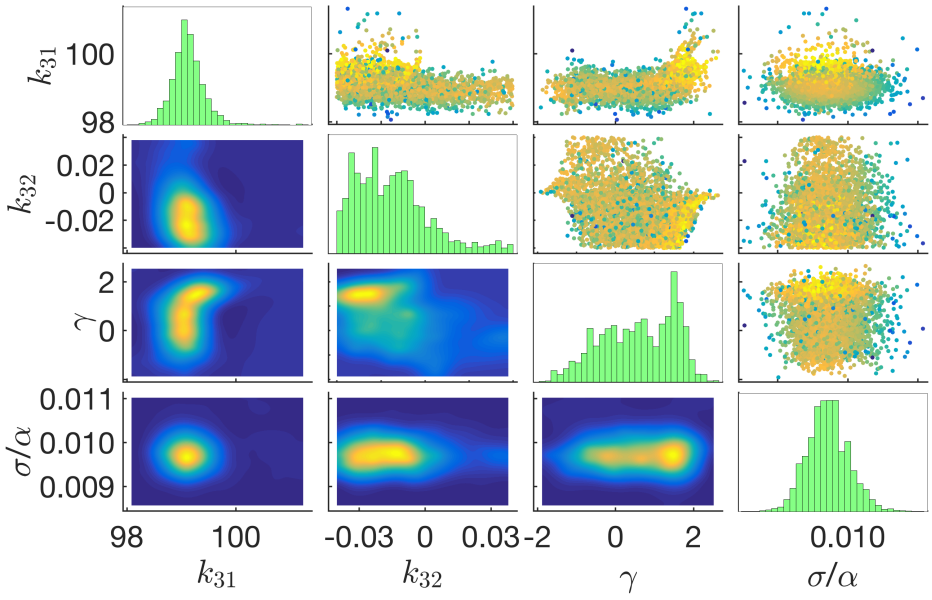
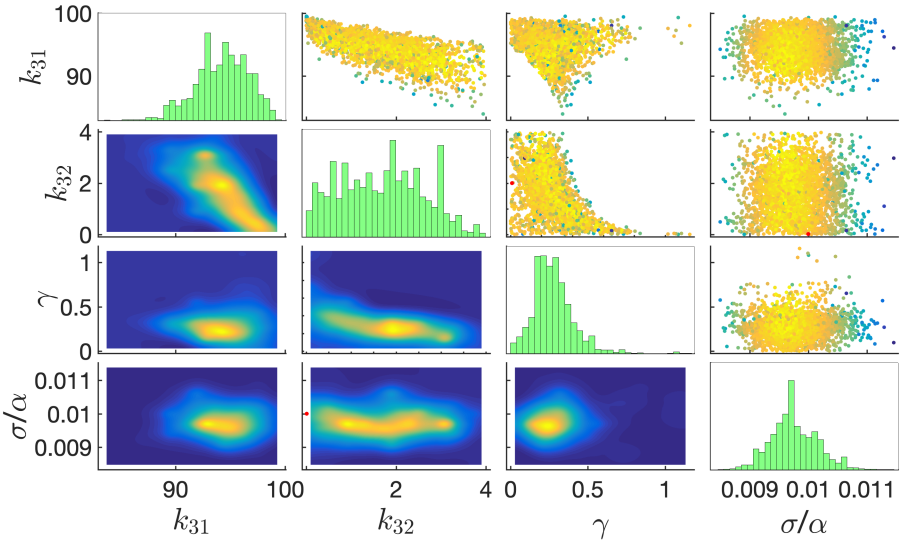


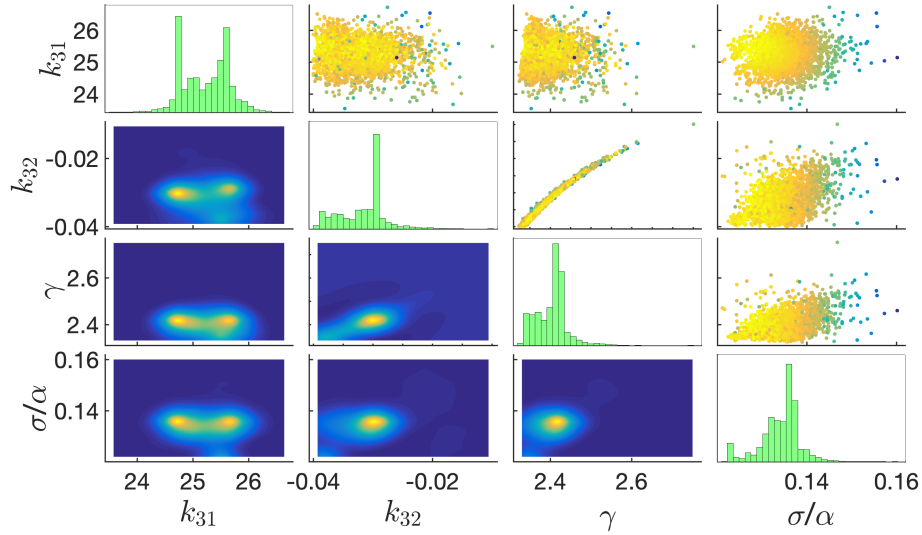
Figure 4: Parameter estimation results using reference data from model where $k_3 = k_{31} + k_{32}x_2^\gamma$ with a time history from $T = 0$ to $T = 10$. The nominal parameter values used were $k_{31} = 100$, $k_{32} = -0.01$, $\gamma = 2$, and noise level $\sigma = 0.01\alpha$. The model used in parameter estimation is $k_3 = k_{31} + k_{32}x_2^\gamma$.Descriptions of the sub-figures can be found in Figure 3.

takes on parameter values in a much larger range and $k_3(x_1)$ is increasing, the prior distributions were expanded to $[0, 400] \times [0, 400] \times [0, 4] \times [0, 0.20]$ and $[-4, 4] \times [-4, 4] \times [-4, 4] \times [0, 0.20]$ for models $k_3(x_1)$ and $k_3(x_2)$, respectively.

To analyze the various models, the results from the previous four experiments are displayed in Table 1. By comparing the Bayes factors for both models that used reference data from $k_3(x_1)$, we find that with probability one that data came from model $k_3(x_1)$, regardless of the range of parameter values that are explored. Correspondingly, model $k_3(x_2)$ recovered parameters that are very different from the nominal parameter values, further demonstrating the inability of that model to recover dynamics similar to those of the observed data. When comparing the Bayes factors for the observation data from $k_3(x_2)$, we find with probability ~ 0.91 that the data came from model $k_3(x_2)$ when we use a sufficiently large prior distribution for model $k_3(x_1)$. However, to achieve this result requires an immense parameter domain that seems infeasible in general as the recovered parameters are 200 times the expected value. Thus, we are able to confidently resolve a model



(a) Results for using model $k_3(x_1)$ on reference data from $k_3(x_2)$.



(b) Results for using model $k_3(x_2)$ on reference data from $k_3(x_1)$.

Figure 5: Parameter estimation results using the "incorrect" model on reference data, i.e. the model $k_3(x_1)$ on noisy data generated from $k_3(x_2)$ and similarly for model $k_3(x_2)$. Descriptions of the sub-figures can be found in Figure 3, and the parameter values used can be found in Figures 3 and 4.

misspecification problem; i.e., we are able to correctly recover the model that generated each of the data sets used.

Model	$p(M_j\|D)$	$\theta_{k_{31}}$	$u_{k_{31}}$ (%)	$\theta_{k_{32}}$	$u_{k_{32}}$ (%)
M_1, O_1	~ 1.0	0.961	37.96	1.045	19.79
M_2, O_1	~ 0.0	0.252	1.64	-3.161	11.89
M_1, O_2	0.0936	94.242	2.52	170.437	5.18
M_2, O_2	0.9064	0.991	0.32	-1.445	-116.22

Model	θ_γ	u_γ (%)	σ/α	$u_{\sigma/\alpha}$ (%)
M_1, O_1	0.999	2.223	0.010	4.22
M_2, O_1	1.202	1.50	0.134	3.24
M_1, O_2	0.138	5.73	0.010	3.94
M_2, O_2	0.288	162.73	0.010	3.08

Table 1: Subset of model selection results for DNA methylation models.

3.3. Two Standard Deviation Added-Noise and Nonlinear Model

Another method to add noise is to use a different standard deviation for each model output. Previously, we had assumed that the noise was added with a constant factor that was the same for x_1, x_2, and x_3. However, the spread of these three populations may be different so we instead considered the case where we had a different standard deviation, σ_1 and σ_3, for x_1 and x_3. The covariance matrix was still assumed to be diagonal, but now $\Sigma_{j,j} = \sigma_1$ for $j = 1, 3, \ldots m-1$ and $\Sigma_{j,j} = \sigma_3$ for $j = 2, 4, \ldots, m$. The noise was added as before, but where we used σ_1 when the output corresponds to an x_1 observation, and σ_3 otherwise, i.e.

$$D_k = \begin{cases} \xi_k + \sigma_1 \epsilon_k, & k = 1, 3, \ldots, m-1, \\ \xi_k + \sigma_3 \epsilon_k, & k = 2, 4, \ldots, m. \end{cases}$$

In the following experiments, we worked with the nonlinear coefficient cases and considered two potential models: $M_{1,3}$, where the coefficients for k_1 and k_3 depend on x_1 and k_2 and k_4 depend on x_3 and M_2, in which all coefficients depend on x_2. For each of these models, we generated reference data as described above using noisy observations from model $M_{1,3}$, denoted $O_{1,3}$, or noisy observations from model M_2, denoted O_2. We first estimated parameters for each model using the noisy data that was generated from

that model. Secondly, we tried to identify which model generated what data. Three experiments were performed: first we attempted to recover the parameters for models $M_{1,3}$ and M_2. Second, model selection was performed to see if the Π4U framework was able to recover which model generated which data set. Finally, we tried to recover all eight model parameters, as well as the corresponding noise levels.

Again, for the ease of comparison, numerical results were computed in terms of the rescaled parameters ($\theta_{k_{11}}$, $\theta_{k_{21}}$, $\theta_{k_{32}}$, $\theta_{k_{41}}$, θ_γ, σ_1/α_1, σ_3/α_3), given by $\theta_{k_{11}} = k_{11}/k_{11_0}$, i.e. the ratio between the estimated value and the nominal value. The prior was assumed uniform on $[0,4]\times[0,4]\times[0,4]\times[0,4]\times[0,0.05]\times[0,0.05]$ in the scaled parameter space. We took the following parameter values for the nominal parameter values $k_{11_0} = 2.1$, $k_{21_0} = 10$, $k_{32_0} = 0.0099$, and $k_{41_0} = 4$ for $M_{1,3}$ and $k_{11_0} = 1.9$, $k_{21_0} = 110$, $k_{32_0} = -0.0099$, and $k_{41_0} = 2$ for M_2.

3.3.1. The Four Most Sensitive Parameters

In this experiment, we used model $M_{1,3}$ with noisy observations $O_{1,3}$ or M_2 with noisy observations O_2 to perform parameter estimation for the four model parameters k_{11}, k_{21}, k_{32}, and k_{41}, which were found to be the four parameters that caused the largest changes in x_1, x_2, and x_3 [19]. In addition, we also tried to recover the amount of added noise. In these experiments, 5% noise was added to the model outputs of x_1 and x_3 to generate observations $O_{1,3}$, where $\sigma_i = 0.05\alpha_i$, for $i = 1, 3$ and α_i is the standard deviation for model outputs x_i.

The case where observations were created using model $M_{1,3}$ has results displayed in Figure 6, where the parameters used for generating the reference data are marked with red dots. We see clear correlations between the parameters: k_{11} and k_{32} have a positive correlation, which makes intuitive sense as those two coefficients have opposite effects: one controls the methylation rate, while the other influences the demethylation rate. Correspondingly, we see positive correlations with k_{21} and k_{41}. In addition to the marginal posterior distributions matching the expected correlations, they also include the nominal parameter values used to create the reference data: for all four model parameters, the "true" parameter value is recovered within one standard deviation; the level of the noise is also recovered well for this case. The marginal distributions from σ_1/α_1 and σ_3/α_3 are both centered around 0.05, the noise level used to create the noisy observations.

Our second case, where model M_2 was used to create noisy observations

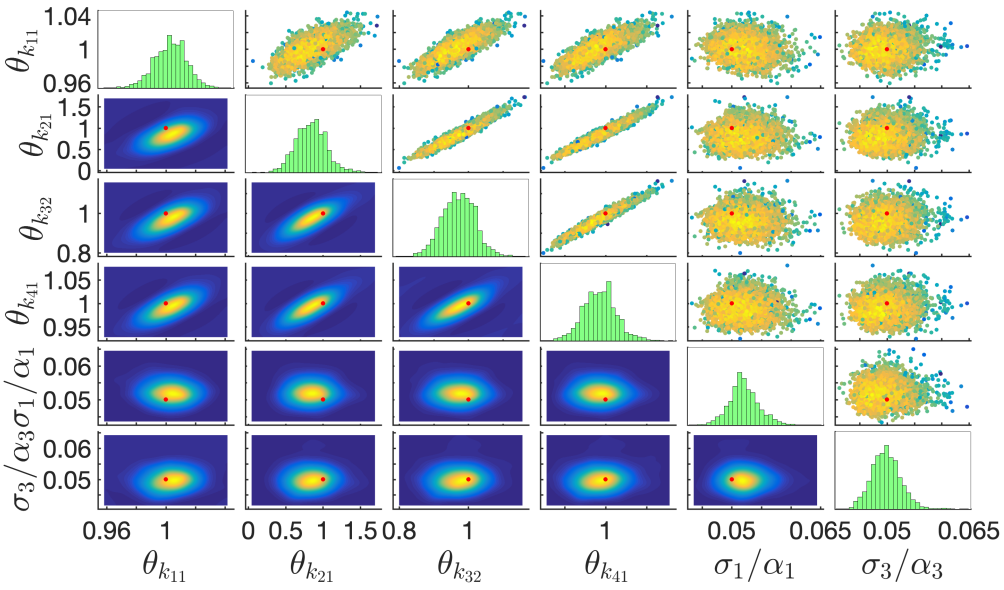


Figure 6: Parameter estimation results for model $M_{1,3}$ using reference data $O_{1,3}$ with 5% added noise. Histograms for each parameter are displayed along the main diagonal of the figure. Sub-figures below the diagonal show the marginal joint density functions for each pair of parameters, while sub-figures above the diagonal show the samples used in the final stage of TMCMC. Colors correspond to probabilities, with yellow likely and blue unlikely. Red dots indicate the parameter values used for generating the noisy data set used for performing parameter estimation and model selection.

O_2, also performs well: as with the previous case, all of the nominal parameter values are recovered within one standard deviation of the estimated means. Furthermore, there are similar correlations to the previous case. Now the correlation between k_{21} and k_{32} is negative due to k_{32} having a negative coefficient: for larger values of k_{21}, a more negative coefficient is needed to have similar dynamics as the observations. In both cases, the relative uncertainty of all parameters are on the same scale. The exception is k_{21} for $M_{1,3}$, which has a relatively bigger uncertainty than the other parameters. This large uncertainty in k_{21} could be due to the relatively larger parameter values that $\theta_{k_{21}}$ explores during the TMCMC algorithm.

Next, we performed model selection. To do this, we used model M_2 on $O_{1,3}$ and model $M_{1,3}$ on O_2 to see whether it could still produce the same dynamic behavior as the other model. We compared both models' results on each set of the noisy observations. In both cases, as seen in Table 2,

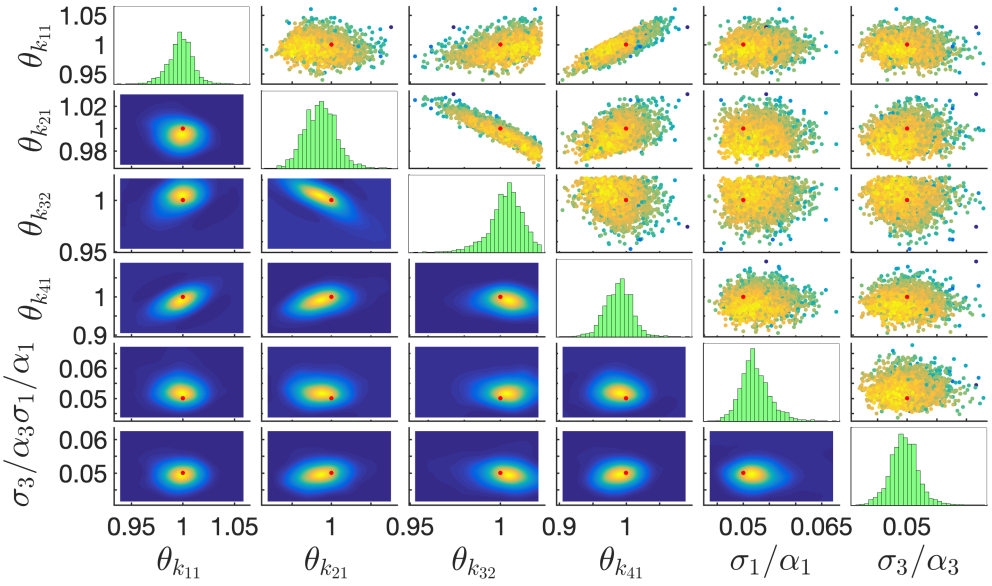


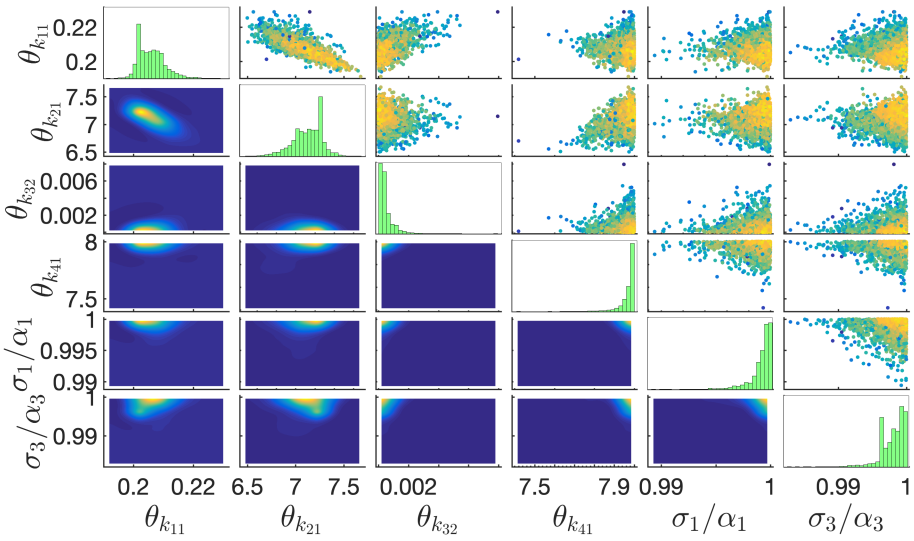
Figure 7: Parameter estimation results for model M_2 using reference data O_2 with 5% added noise.

the other model is unable to recover the dynamics using the uniform prior $[0, 8] \times [0, 8] \times [0, 8] \times [0, 8] \times [0, 1] \times [0, 1]$. When the Bayes factors for the two models are compared, the likelihood of M_2 on $O_{1,3}$ is negligible, and similarly for M_1 on O_2. These results can be intuitively seen by also looking at how the incorrect models are inadequate at estimating the parameters, as they are not able to match the dynamics of the problem. Demonstrated in Figure 8, the "best" parameters in these cases are along the border of the prior distribution and to better match the observed data, portraying that parameter values much larger than those expected biologically are required to recover dynamics close to the noisy observations.

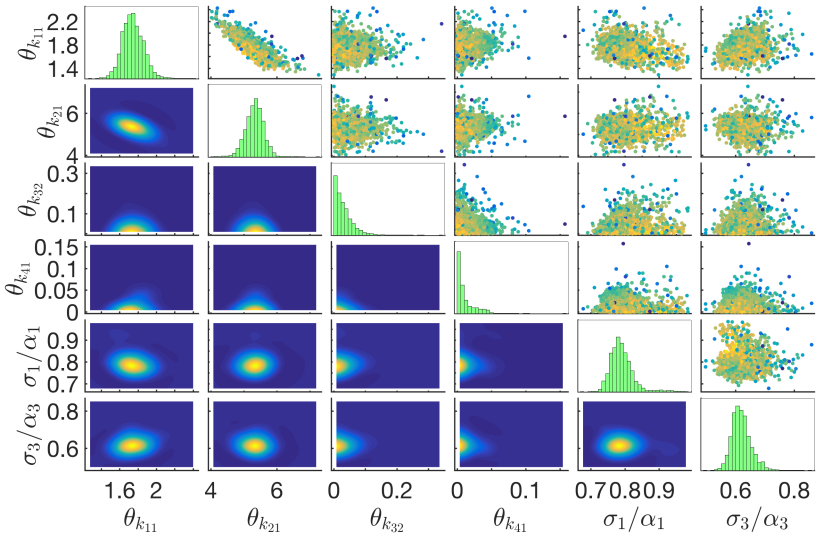
3.4. All Parameters

Finally, we attempted to recover the eight coefficients for all four parameters k_1, k_2, k_3, and k_4 using noisy observations from model $M_{1,3}$ and M_2. We did this for two cases of added noise: 1% and 5% noise.

For the 1% added noise case using model $M_{1,3}$ displayed in Figure 9, all parameters are recovered within two standard deviations of the means. As in the four parameter case, the distribution for k_{21} has a rather wide



(a) Results for using model $M_{1,3}$ on reference data from O_2.



(b) Results for using model M_2 on reference data from $O_{1,3}$.

Figure 8: Parameter estimation results using a different model than the one that generated the reference data. Descriptions of the sub-figures can be found in Figure 3, and the parameter values used can be found in Figures 6 and 7.

Model	$p(M_i\|D)$	$\theta_{k_{11}}$	$u_{k_{11}}$ (%)	$\theta_{k_{21}}$	$u_{k_{21}}$ (%)	$\theta_{k_{32}}$
$M_{1,3}$, $O_{1,3}$	~ 1.00	1.0030	1.09	0.8374	23.93	0.9710
M_2, $O_{1,3}$	~ 0.00	1.7466	6.66	5.3280	5.21	0.0316
$M_{1,3}$, O_2	~ 0.00	0.2060	2.14	7.1298	2.09	0.0004
M_2, O_2	~ 1.00	0.9975	1.20	0.9945	0.80	1.0032

Model	$u_{k_{32}}$ (%)	$\theta_{k_{41}}$	$u_{k_{41}}$ (%)	σ_1/α_1	u_{σ_1/α_1}	σ_3/α_3	u_{σ_3/α_3}
$M_{1,3}$, $O_{1,3}$	4.82	0.9926	2.01	0.0521	4.78	0.0498	5.36
M_2, $O_{1,3}$	92.75	0.0112	112.73	0.7903	4.37	0.6196	5.34
$M_{1,3}$, O_2	115.23	7.9701	0.47	0.9992	0.10	0.9980	0.17
M_2, O_2	0.92	0.9884	1.99	0.0524	5.39	0.0495	4.77

Table 2: Model selection results for estimating four most sensitive transition rates for two model scenarios and two observed data sets for 5% added noise.

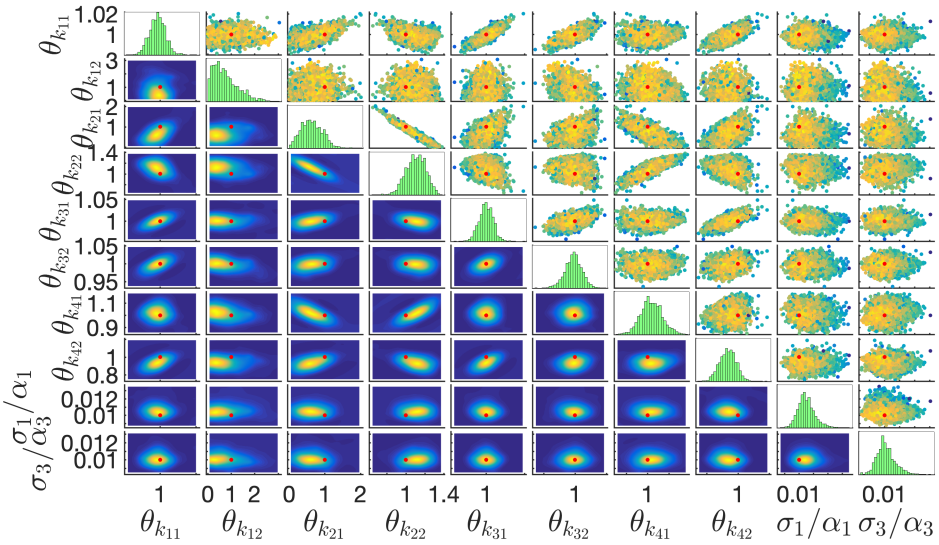


Figure 9: Parameter estimation results for model $M_{1,3}$ using reference data $O_{1,3}$ with 1% added noise.

standard deviation. Furthermore, we note that the parameters k_{12}, k_{22}, k_{42} all have wider distributions than the more sensitive parameters k_{11}, k_{32}, and k_{41}. This matches intuition, as the parameters that cause larger changes in the model outputs are more easily recovered since we want to minimize the difference between model outputs and observed data. In addition, the model

correlations match what is expected from the system dynamics: for example, k_{21} and k_{22} have a negative correlation. Since the coefficient $k_2(x_3)$ is an increasing function, a larger value of k_{21} would require a smaller value of k_{22} to result in similar values for $k_2(x_3)$.

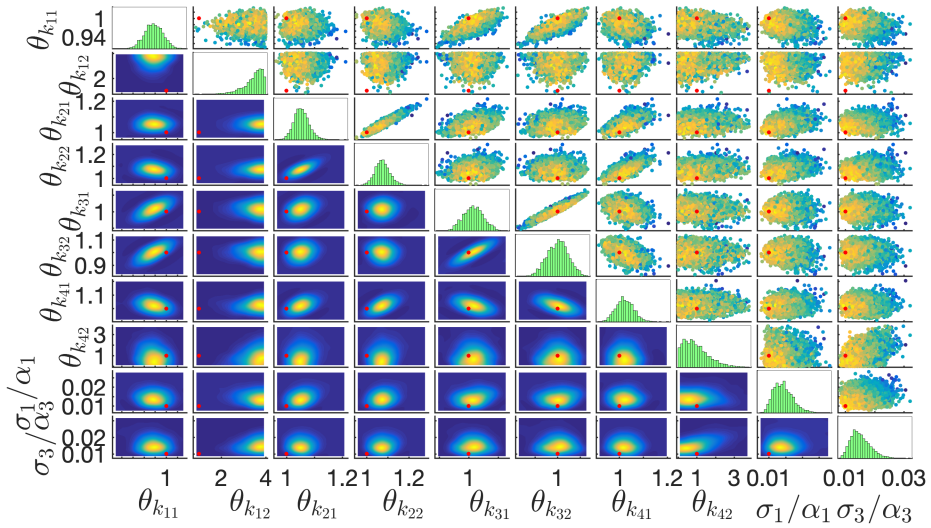


Figure 10: Parameter estimation results for model M_2 using reference data O_2 with 1% added noise.

In addition, we again considered the 1% added noise case for M_2, shown in Figure 10. For the model where coefficients depend on x_2 all nominal parameter values are found within two standard deviations of the recovered means, except k_{12}. It was found that, model M_2 is highly insensitive to k_{12}: increasing or decreasing k_{12} by 50% change the model outputs on the order of 10^{-5}; however, the added noise adjusts the parameter values on the order of 10^{-1}. Due to this discrepancy in the model sensitivity and noise level, it seems that the parameter value for k_{12} is difficult to recover and, as a result of model robustness to this parameter, is not as important to recover accurately. Similar to the $M_{1,3}$ case, the recovered marginal distributions have correlations that match intuition. For example, k_{21} and k_{22} have a positive correlation. Since $k_2(x_2)$ is a decreasing function, smaller values of k_{21} correspondingly need smaller values of k_{22} to keep k_2 close to the same values as the nominal parameter values.

Model selection was also performed on these parameter values. The scaled

Model	$p(M_j\|D)$	$\theta_{k_{11}}$	$u_{k_{11}}$ (%)	$\theta_{k_{12}}$	$u_{k_{12}}$ (%)	$\theta_{k_{21}}$	$u_{k_{21}}$ (%)
$M_{1,3}$, $O_{1,3}$	~ 1.0	0.999	0.43	0.743	71.00	0.636	46.81
M_2, $O_{1,3}$	~ 0.0	3.899	0.85	0.650	59.11	4.000	0.01
$M_{1,3}$, O_2	~ 0.0	0.326	1.14	0.011	118.97	2.943	6.43
M_2, O_2	~ 1.0	0.992	0.52	3.151	16.66	1.010	0.88

$\theta_{k_{22}}$	$u_{k_{22}}$	$\theta_{k_{31}}$	$u_{k_{31}}$	$\theta_{k_{32}}$	$u_{k_{32}}$	$\theta_{k_{41}}$	$u_{k_{42}}$
1.106	9.36	1.000	0.90	0.998	1.34	1.018	4.88
0.006	115.00	1.740	0.84	0.001	99.19	0.000	92.20
0.001	102.08	4.000	0.05	0.001	114.44	1.400	7.75
1.017	1.13	1.000	0.80	0.996	1.16	1.003	0.99

$\theta_{k_{42}}$	$u_{k_{42}}$	σ_1/α_1	u_{σ_1/α_1} (%)	σ_3/α_3	u_{σ_3/α_3} (%)
0.934	5.48	0.011	5.57	0.010	5.90
3.696	6.05	0.200	0.0003	0.200	0.01
0.005	94.86	0.200	0.03	0.200	0.02
0.536	72.47	0.01	6.18	0.01	6.37

Table 3: Model selection results for estimating all transition rates added noise for two model scenarios and two observed data sets for 1% added noise.

model parameters were given a prior of $[0, 4]$ and the standard deviations had a uniform prior on $[0, 0.20]$. In these cases, we find that $M_{1,3}$ is the model that best matches $O_{1,3}$ with probability one, and similarly M_2 best matches O_2 with probability one. In addition, the incorrect models are unable to recover the proper parameter values. As seen in Table 3, many of the recovered parameter values are on the boundary of the prior distribution, demonstrating that the usual domain for the parameters of each model are not able to recover the dynamics of the other model. In both of our cases, we are able to recover that $O_{1,3}$ came from $M_{1,3}$ and O_2 is a noisy version of model M_2.

The same experiment was repeated using 5% added noise. Using model $M_{1,3}$ (seen in Figure 11), the 5% noise case recovers all parameters within two standard deviations, although the distributions are comparatively wider than those found in previous experiments: most of the recovered standard deviations are 2–5 times as large for 5% added noise as the 1% added noise case. Again, this matches intuition as noisier data should result in more uncertainty in the estimated distributions.

For model M_2, more model parameters are not estimated well: k_{12}, k_{31},

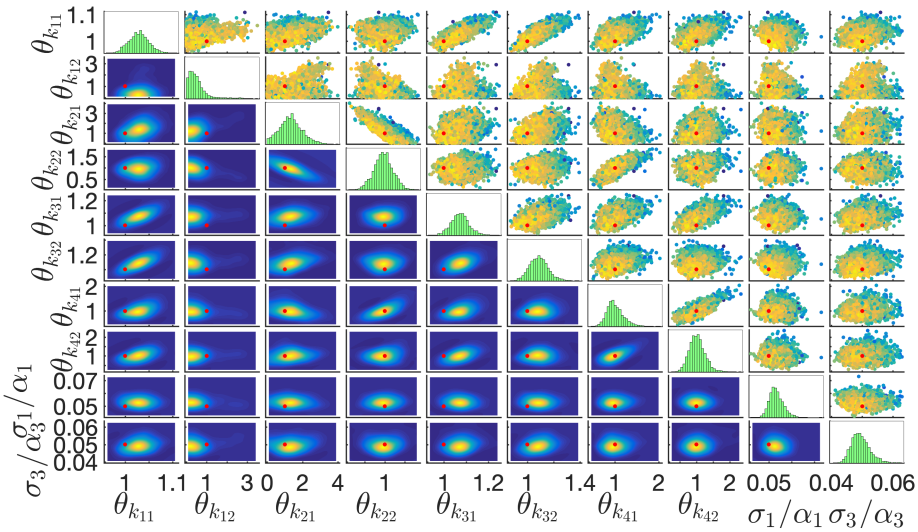


Figure 11: Parameter estimation results $M_{1,3}$ using reference data $O_{1,3}$ with 5% added noise.

k_{32}, and k_{42} are not recovered within two standard deviations, and most parameters are only recovered due to having very large standard deviations. For M_2, all standard deviations for the 5% noise case are 2–5 times as large as those for 1% noise.

Finally, the model selection was again performed and similar results to the 1% noise case occurred: again, the Bayes factors accurately conclude which observation data came from which model. Secondly, the distributions again move to the boundaries of the domain, demonstrating that for the typical parameter values, M_2 cannot give the dynamics of $M_{1,3}$ and vice versa.

4. Conclusions

In this paper, we presented an uncertainty quantification framework that is applicable to a wide array of parameter estimation and model selection problems. For experimental biologists with wet lab data, the methodology can be used to test and improve various models, as well as guide future research. In order to use the method described in the paper, three components are necessary: first, experimental data are needed as the reference data. Second, model or models for the biological phenomena are needed to estimate parameters or test which model best describes the experimental data.

Model	$p(M_j\|D)$	$\theta_{k_{11}}$	$u_{k_{11}}$ (%)	$\theta_{k_{12}}$	$u_{k_{12}}$ (%)	$\theta_{k_{21}}$	$u_{k_{21}}$ (%)
$M_{1,3}$, $O_{1,3}$	~ 1.0	1.029	2.01	0.521	88.73	1.34	47.35
M_2, $O_{1,3}$	~ 0.0	3.429	9.10	1.269	68.52	3.907	3.08
$M_{1,3}$, O_2	~ 0.0	0.340	3.36	0.078	100.10	3.687	4.93
M_2, O_2	~ 1.0	0.925	2.23	0.667	69.54	1.173	4.78

$\theta_{k_{22}}$	$u_{k_{22}}$	$\theta_{k_{31}}$	$u_{k_{31}}$	$\theta_{k_{32}}$	$u_{k_{32}}$	$\theta_{k_{41}}$	$u_{k_{42}}$
0.967	23.81	1.073	3.75	1.087	6.50	1.008	18.74
0.334	90.12	1.559	10.10	0.126	102.09	0.007	106.47
0.004	81.31	3.986	0.34	0.004	109.10	1.869	6.33
1.244	5.53	0.995	5.46	0.917	7.37	1.139	4.96

$\theta_{k_{42}}$	$u_{k_{42}}$	σ_1/α_1	u_{σ_1/α_1} (%)	σ_3/α_3	u_{σ_3/α_3} (%)
1.017	23.21	0.053	5.61	0.050	6.10
1.765	59.56	0.837	5.05	0.665	5.80
0.031	102.78	0.499	0.18	0.499	0.12
1.061	81.26	0.55	6.91	0.051	6.87

Table 4: Model selection results for estimating all transition rates added noise for two model scenarios and two observed data sets for 5% added noise.

Finally, a "connector" code is required to compare the model to the experimental data. Our Bayesian framework propagates the forward model through sampled sets of parameters and uses the goodness of the fit to guide the next generation of samples. In addition, this approach is highly parallelizable, lending itself easily to applications that were previously too computationally intensive to be feasible [30]. Due to these simple three requirements and its parallelizability, this method is useable for parameter estimation and model selection problems for a wide array of fields and backgrounds.

We find that the Bayesian uncertainty quantification framework allows us to recover robust predictions for parameters of the DNA methylation model, for example the the demethylation rate k_3, despite only using a limited amount of noisy data. When examining the model with unknown exponent γ, the method was able to accurately recover all parameters within two standard deviations of the nominal parameter values for observed data that was generated from the corresponding model. Furthermore, the method was able to discern which model the observed data were generated from, as long as the biological constraints were maintained. Our methodology prefers that demethylation is dependent upon the total number of unmethylated CpG

dyads for the noisy data generated from that assumption. If the parameter space searched allows the model to no longer obey the observed monotonicity of the demethylation parameters, the observed data generated from the demethylation parameter depending on hemimethylated CpG dyads x_2, i.e. $k_3(x_2)$, can be matched to both models.

In our second scenario, we considered various noise levels and estimated more model parameters for the demethylation rates, although we assumed that the exponent was fixed $\gamma = 2$ for all populations. For the four most sensitive parameter scenario, all parameters were recovered within a reasonable deviation from the nominal values. Furthermore, correlations that match the intuition of the biological model are observed in the various recovered parameter values. Finally, we were also able to identify whether the parameters depended on the unmethylated and methylated or the hemimethylated CpG dyads, showing that our method may be able to help better understand the biological process if given real data.

Finally, we estimate all eight parameters related to demethylation rates for a few added noise scenarios to demonstrate the power of the Bayesian framework. Again, when the parameters depend on unmethylated and methylated CpG dyads, all values are recovered within two standard deviations of the nominal parameter values. The scenario where the parameters depend on the hemimethylated CpG dyads has a bit more difficulty recovering the nominal parameter values for 5% added noise, but this is likely due to some model insensitivity to those parameter values. In all cases, however, we are able to recover which model generated what observed data. In addition, by comparing the recovered levels of noise, we observe that the $\sigma = 0.05\alpha$ scenarios correspondingly have larger uncertainties in the recovered parameter values than the $\sigma = 0.01\alpha$ cases.

It is important to put the results of our work within a biological context. For example, a finding of the work is the ability of our technique to determine if model parameter values are dependent on different methylation states. The issue of whether or not rates of methylation/demethylation are inexorably linked to promoter topology is an ongoing question the experimental community has been attempting to unravel. For instance, it is generally regarded that during replication DNMTs are responsible for the remethylation of hemimethylated DNA. However, due to a number of experimental findings it has been suggested by [33] that this is unfeasible and that the methylation of a CpG site is affected by the methylation levels of the nearby CpG sites. This idea couples DNMTs activity with CpG level. Therefore,

the fact that our theoretical analysis can explore this issue is a significant feature of our work which can be explored further by the experimental community. Another area of biological significance is that our models are based on the assumption of enzymatic processivity [34]. A way to fully test the findings from our model would be determine in the laboratory whether our findings hold in case it turns out that DNMTs are not processive enzymes and that rather other factors are at play, as for example, RNA directed DNA methylation.

Acknowledgements

Part of this research was conducted using computational resources and services at the Center for Computation and Visualization, Brown University. **Funding**: NK acknowledges financial support by the International Research Excellence Awards 2017/18, University of Chester, under the research project "Computational modelling and stochastic analysis of DNA dynamics." NK would also like to thank the Division of Applied Mathematics at Brown University for their hospitality during his visit, which was funded by a divisional IBM fund. LZ would like to acknowledge the University of Chester for funding his PhD studentship. KL and AM were partially supported by the NSF through grants DMS-1521266 and DMS-1552903. The authors declare no competing interests.

[1] Z. D. Smith, A. Meissner, DNA methylation: roles in mammalian development, Nature Reviews Genetics 14 (2013) 204–220. doi:10.1038/nrg3354.

[2] M. J. Jones, S. J. Goodman, M. S. Kobor, DNA methylation and healthy human aging, Aging Cell 14 (6) (2015) 924–932. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/acel.12349, doi:10.1111/acel.12349.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/acel.12349

[3] M. Klutstein, D. Nejman, R. Greenfield, H. Cedar, DNA methylation in cancer and aging, Cancer Research 76 (12) (2016) 3446–3450. arXiv:http://cancerres.aacrjournals.org/content/76/12/3446.full.pdf, doi:10.1158/0008-5472.CAN-15-3278.
URL http://cancerres.aacrjournals.org/content/76/12/3446

[4] J. Zhong, G. Agha, A. A. Baccarelli, The role of DNA methylation in cardiovascular risk and disease, Circulation Research 118 (1) (2016) 119–131. arXiv:http://circres.ahajournals.org/content/118/1/119.full.pdf, doi:10.1161/CIRCRESAHA.115.305206.
URL http://circres.ahajournals.org/content/118/1/119

[5] P. L. De Jager, G. Srivastava, K. Lunnon, J. Burgess, L. C. Schalkwyk, L. Yu, M. L. Eaton, B. T. Keenan, J. Ernst, C. McCabe, A. Tang, T. Raj, J. Replogle, W. Brodeur, S. Gabriel, H. S. Chai, C. Younkin, S. G. Younkin, F. Zou, M. Szyf, C. B. Epstein, J. A. Schneider, B. E. Bernstein, A. Meissner, N. Ertekin-Taner, L. B. Chibnik, M. Kellis, J. Mill, D. A. Bennett, Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci, Nature Neuroscience 17 (2014) 1156–1163. doi:10.1038/nn.3786.
URL http://dx.doi.org/10.1038/nn.3786

[6] J. Delgado-Calle, A. F. Fernández, J. Sainz, M. T. Zarrabeitia, C. Sañudo, R. García-Renedo, M. I. Pérez-Núñez, C. García-Ibarbia, M. F. Fraga, J. A. Riancho, Genome-wide profiling of bone reveals differentially methylated regions in osteoporosis and osteoarthritis, Arthritis & Rheumatism 65 (1) (2013) 197–205. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/art.37753, doi:10.1002/art.37753.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/art.37753

[7] S. Horvath, DNA methylation age of human tissues and cell types, Genome Biology 14 (10) (2013) 3156. doi:10.1186/gb-2013-14-10-r115.
URL https://doi.org/10.1186/gb-2013-14-10-r115

[8] S. Horvath, M. Gurven, M. E. Levine, B. C. Trumble, H. Kaplan, H. Allayee, B. R. Ritz, B. Chen, A. T. Lu, T. M. Rickabaugh, B. D. Jamieson, D. Sun, S. Li, W. Chen, L. Quintana-Murci, M. Fagny, M. S. Kobor, P. S. Tsao, A. P. Reiner, K. L. Edlefsen, D. Absher, T. L. Assimes, An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease, Genome Biology 17 (1) (2016) 171. doi:10.1186/s13059-016-1030-0.
URL https://doi.org/10.1186/s13059-016-1030-0

[9] P. A. Jones, G. Liang, Rethinking how DNA methylation patterns are maintained, Nature Reviews Genetics 10 (2009) 805–811. doi:10.1038/nrg2651.

[10] M. Esteller, J. M. Silva, G. Dominguez, F. Bonilla, X. Matias-Guiu, E. Lerma, E. Bussaglia, J. Prat, I. C. Harkes, E. A. Repasky, E. Gabrielson, M. Schutte, S. B. Baylin, J. G. Herman, Promoter hypermethylation and brca1 inactivation in sporadic breast and ovarian tumors, JNCI: Journal of the National Cancer Institute 92 (7) (2000) 564–569. doi:10.1093/jnci/92.7.564.
URL http://dx.doi.org/10.1093/jnci/92.7.564

[11] A. Hunter, P. A. Spechler, A. Cwanger, Y. Song, Z. Zhang, G.-s. Ying, A. K. Hunter, E. deZoeten, J. L. Dunaief, DNA methylation is associated with altered gene expression in AMD, Investigative Ophthalmology & Visual Science 53 (4) (2012) 2089. doi:10.1167/iovs.11-8449.

[12] K. D. Robertson, DNA methylation, methyltransferases, and cancer, Oncogene 20 (2001) 3139–3155. doi:10.1038/sj.onc.1204341.

[13] K. S. Crider, T. P. Yang, R. J. Berry, L. B. Bailey, Folate and DNA methylation: A review of molecular mechanisms and the evidence for folate's role, Advances in Nutrition 3 (1) (2012) 21–38. doi:10.3945/an.111.000992.
URL http://dx.doi.org/10.3945/an.111.000992

[14] K. D. Robertson, E. Uzvolgyi, G. Liang, C. Talmadge, J. Sumegi, F. A. Gonzales, P. A. Jones, The human DNA methyltransferases (DNMTs) 1, 3a and 3b: coordinate mRNA expression in normal tissues and overexpression in tumors, Nucleic Acids Research 27 (11) (1999) 2291–2298. doi:10.1093/nar/27.11.2291.
URL http://dx.doi.org/10.1093/nar/27.11.2291

[15] Z.-x. Chen, A. D. Riggs, DNA methylation and demethylation in mammals, Journal of Biological Chemistry 286 (21) (2011) 18347–18353. arXiv:http://www.jbc.org/content/286/21/18347.full.pdf+html, doi:10.1074/jbc.R110.205286.
URL http://www.jbc.org/content/286/21/18347.abstract

[16] A. Razin, A. Riggs, DNA methylation and gene function, Science 210 (4470) (1980) 604–610. arXiv:http://science.sciencemag.org/content/210/4470/604.full.pdf, doi:10.1126/science.6254144.
URL http://science.sciencemag.org/content/210/4470/604

[17] L. Scourzic, E. Mouly, O. A. Bernard, TET proteins and the control of cytosine demethylation in cancer, Genome Medicine 7 (1) (2015) 9. doi:10.1186/s13073-015-0134-6.
URL https://doi.org/10.1186/s13073-015-0134-6

[18] M. T. Mc Auley, K. M. Mooney, J. E. Salcedo-Sora, Computational modelling folate metabolism and DNA methylation: implications for understanding health and ageing, Briefings in Bioinformatics 19 (2) (2018) 303–317. doi:10.1093/bib/bbw116.
URL http://dx.doi.org/10.1093/bib/bbw116

[19] L. Zagkos, M. M. Auley, J. Roberts, N. I. Kavallaris, Mathematical models of DNA methylation dynamics: Implications for health and ageing, Journal of Theoretical Biology 462 (2019) 184 – 193. doi:https://doi.org/10.1016/j.jtbi.2018.11.006.

[20] A. P. McGovern, B. E. Powell, T. J. Chevassut, A dynamic multicompartmental model of DNA methylation with demonstrable predictive value in hematological malignancies, Journal of Theoretical Biology 310 (2012) 14 – 20. doi:https://doi.org/10.1016/j.jtbi.2012.06.018.
URL http://www.sciencedirect.com/science/article/pii/S0022519312003050

[21] A. Jeltsch, R. Z. Jurkowska, New concepts in DNA methylation, Trends in Biochemical Sciences 39 (7) (2014) 310 – 318. doi:https://doi.org/10.1016/j.tibs.2014.05.002.
URL http://www.sciencedirect.com/science/article/pii/S0968000414000875

[22] R. A. Waterland, K. B. Michels, Epigenetic epidemiology of the developmental origins hypothesis, Annual Review of Nutrition 27 (1) (2007) 363–388, pMID: 17465856. arXiv:https://doi.org/10.1146/annurev.nutr.27.061406.093705, doi:10.1146/annurev.nutr.27.061406.093705.
URL https://doi.org/10.1146/annurev.nutr.27.061406.093705

[23] K. Lokk, V. Modhukur, B. Rajashekar, K. Märtens, R. Mägi, R. Kolde, M. Koltšina, T. K. Nilsson, J. Vilo, A. Salumets, N. Tõnisson, DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns, Genome Biology 15 (4) (2014) 3248. doi:10.1186/gb-2014-15-4-r54.
URL https://doi.org/10.1186/gb-2014-15-4-r54

[24] J. O. Haerter, C. Lövkvist, I. B. Dodd, K. Sneppen, Collaboration between cpg sites is needed for stable somatic inheritance of DNA methylation states, Nucleic Acids Research 42 (4) (2014) 2235–2244. doi:10.1093/nar/gkt1235.
URL http://dx.doi.org/10.1093/nar/gkt1235

[25] C. Lövkvist, I. B. Dodd, K. Sneppen, J. O. Haerter, DNA methylation in human epigenomes depends on local topology of cpg sites, Nucleic Acids Research 44 (11) (2016) 5123–5132. doi:10.1093/nar/gkw124.
URL http://dx.doi.org/10.1093/nar/gkw124

[26] J. L. Beck, K.-V. Yuen, Model selection using response measurements: Bayesian probabilistic approach, Journal of Engineering Mechanics 130 (2) (2004) 192–203.

[27] K.-V. Yuen, Bayesian methods for structural dynamics and civil engineering, John Wiley & Sons, 2010.

[28] P. E. Hadjidoukas, P. Angelikopoulos, C. Papadimitriou, P. Koumoutsakos, Π4U: A high performance computing framework for bayesian uncertainty quantification of complex models, Journal of Computational Physics 284 (2015) 1–21.

[29] K. Larson, C. Bowman, Z. Chen, P. Hadjidoukas, C. Papadimitriou, P. Koumoutsakos, A. Matzavinos, Data-driven prediction and origin identification of epidemics in population networks, *Submitted*.

[30] C. Bowman, K. Larson, A. Roitershtein, D. Stein, A. Matzavinos, Bayesian uncertainty quantification for particle-based simulation of lipid bilayer membranes, in: M. Stolarska, N. Tarfulea (Eds.), Cell Movement: Modeling and Applications, Springer, 2018, pp. 77–102. doi:10.1007/978-3-319-96842-1_4.

[31] J. Ching, Y.-C. Chen, Transitional markov chain monte carlo method for bayesian model updating, model class selection, and model averaging, Journal of engineering mechanics 133 (7) (2007) 816–832.

[32] D. Barber, Bayesian Reasoning and Machine Learning, Cambridge University Press, 2012.

[33] C. Lövkvist, I. B. Dodd, K. Sneppend, J. O. Haerter, Dna methylation in human epigenomes depends on local topology of cpg sites, Nucleic Acids Research 44 (11) (2016) 5123–5132.

[34] E. Hervouet, P. Peixoto, R. Delage-Mourroux, M. Boyer-Guittaut, P.-F. Cartron, Specific or not specific recruitment of dnmts for dna methylation, an epigenetic dilemma, Clinical Epigenetics 10 (1) (2018) 17. doi:10.1186/s13148-018-0450-y.
URL https://doi.org/10.1186/s13148-018-0450-y